

AU/ACSC/042/2001-04

AIR COMMAND AND STAFF COLLEGE

AIR UNIVERSITY

STATISTICAL ANALYSIS  
OF MULTIPLE CHOICE TESTING

by

Mark A. Colbert, Major, USAF

A Research Report Submitted to the Faculty

In Partial Fulfillment of the Graduation Requirements

Advisor: Lieutenant Colonel Thomas P. Himes, Jr.

Maxwell Air Force Base, Alabama

April 2001

Report Documentation Page		
<b>Report Date</b> 01APR2001	<b>Report Type</b> N/A	<b>Dates Covered (from... to)</b> -
<b>Title and Subtitle</b> Statistical Analysis of Multiple Choice Testing		<b>Contract Number</b>
		<b>Grant Number</b>
		<b>Program Element Number</b>
<b>Author(s)</b> Colbert, Mark A.		<b>Project Number</b>
		<b>Task Number</b>
		<b>Work Unit Number</b>
<b>Performing Organization Name(s) and Address(es)</b> Air Command and Staff College Air University Maxwell AFB, AL		<b>Performing Organization Report Number</b>
<b>Sponsoring/Monitoring Agency Name(s) and Address(es)</b>		<b>Sponsor/Monitor's Acronym(s)</b>
		<b>Sponsor/Monitor's Report Number(s)</b>
<b>Distribution/Availability Statement</b> Approved for public release, distribution unlimited		
<b>Supplementary Notes</b>		
<b>Abstract</b>		
<b>Subject Terms</b>		
<b>Report Classification</b> unclassified	<b>Classification of this page</b> unclassified	
<b>Classification of Abstract</b> unclassified	<b>Limitation of Abstract</b> UU	
<b>Number of Pages</b> 56		

## **Disclaimer**

The views expressed in this academic research paper are those of the author and do not reflect the official policy or position of the US government or the Department of Defense. In accordance with Air Force Instruction 51-303, it is not copyrighted, but is the property of the United States government.

## *Contents*

	<i>Page</i>
DISCLAIMER .....	ii
ILLUSTRATIONS .....	v
PREFACE .....	vi
ABSTRACT .....	vii
BACKGROUND .....	1
Introduction and Problem Definition .....	1
Scope of Analysis .....	2
Thesis .....	2
Methodology .....	2
HISTORICAL REVIEW .....	3
Why Use Statistics for Test Item Analysis .....	3
Definitions .....	4
Ease Index .....	4
Differentiation Index .....	4
Correlation Coefficients .....	5
Quantitative Factors .....	5
Qualitative Factors .....	8
ACSC Distance Learning Department's Current Methods .....	10
ANALYSIS .....	13
TAD Software Program .....	13
Inputs and Controls .....	13
Output Format .....	14
User Friendliness .....	15
ITEMAN Software Program .....	15
Inputs and Controls .....	15
Output Format .....	16
User Friendliness .....	17
Quantitative Analysis .....	17
Qualitative Analysis .....	19
ACSC Distance Learning Department's Current Methods Analysis .....	20

CONCLUSIONS.....	24
Summary of Findings .....	24
ACSC Distance Learning Department's Current Methods .....	24
<i>TAD</i> versus <i>ITEMAN</i> .....	25
Recommendations .....	26
<i>ITEMAN</i> as the Preferred Program .....	26
Quantitative Measurements .....	26
Qualitative Guidelines .....	27
SAMPLE OUTPUTS FROM SOFTWARE PROGRAMS .....	28
<i>TAD</i> Sample Output Using EI, DI, and Point Biserial Correlations.....	28
<i>ITEMAN</i> Sample Output Using EI, DI, and Point Biserial Correlations.....	30
<i>ITEMAN</i> Sample Output Using EI, DI, and Biserial Correlations .....	35
QUESTION WRITING GUIDELINES .....	40
Maxwell Academic Instructor School <i>Test Item Analysis Handout's</i> section on Qualitative Analysis.....	40
James D. Hansen and Lee Dexter's Item-writing Guidelines.....	41
GLOSSARY .....	44
Abbreviations.....	44
Definitions .....	44
BIBLIOGRAPHY .....	48

## *Illustrations*

	<i>Page</i>
Figure 1 Unit Normal Curve and Values of q, p and y .....	8
Figure 2 <i>TAD</i> “Item Statistics” window sample output.....	14
Figure 3 <i>ITEMAN</i> Sample Output Using EI, DI and Biserial Coefficients .....	16
Figure 4 <i>TAD</i> Sample Output using EI, DI and Point Biserial (Part 1 of 2).....	28
Figure 5 <i>TAD</i> Sample Output using EI, DI and Point Biserial (Part 2 of 2).....	29
Figure 6 <i>ITEMAN</i> Sample Output Using EI, DI and Point Biserial (Part 1 of 5).....	30
Figure 7 <i>ITEMAN</i> Sample Output Using EI, DI and Point Biserial (Part 2 of 5).....	31
Figure 8 <i>ITEMAN</i> Sample Output Using EI, DI and Point Biserial (Part 3 of 5).....	32
Figure 9 <i>ITEMAN</i> Sample Output Using EI, DI and Point Biserial (Part 4 of 5).....	33
Figure 10 <i>ITEMAN</i> Sample Output Using EI, DI and Point Biserial (Part 5 of 5).....	34
Figure 11 <i>ITEMAN</i> Sample Output Using EI, DI and Biserial (Part 1 of 5) .....	35
Figure 12 <i>ITEMAN</i> Sample Output Using EI, DI and Biserial (Part 2 of 5) .....	36
Figure 13 <i>ITEMAN</i> Sample Output Using EI, DI and Biserial (Part 3 of 5) .....	37
Figure 14 <i>ITEMAN</i> Sample Output Using EI, DI and Biserial (Part 4 of 5) .....	38
Figure 15 <i>ITEMAN</i> Sample Output Using EI, DI and Biserial (Part 5 of 5) .....	39

## ***Preface***

The United States Air Force Air Command and Staff College's Distance Learning Department offered this research topic as an opportunity to evaluate their process of analyzing multiple-choice questions used in their tests. I selected this topic because I have an interest in the topic and have a bachelor's degree in mathematics. Like many people who have taken multiple-choice question tests, I have always wondered how test givers decide to throw out "bad" questions. This research paper was an excellent opportunity to explore this question in depth.

I would like to acknowledge Lieutenant Colonel Thomas Himes, my research advisor, for his support and background knowledge on the subject. I would like to thank Dr. Thomas R. Renckly and Mr. Michael Zieky for allowing me to interview them at length to gain from their expert knowledge in the test item analysis and evaluation field. I would also like to thank many others too numerous to name who have helped me complete this research effort.

### ***Abstract***

The multiple-choice question exam is a very popular method of evaluation used by educators everywhere. The Air Command and Staff College Distance Learning Department uses multiple-choice exams for testing non-residence students. ACSC currently uses the *Test Analysis and Development (TAD)* software program's two quantitative measurements, Ease Index and Differentiation Index, to flag possible problem questions for qualitative review. They also use student feedback to flag questions for review. ACSC uses the Maxwell Academic Instructor School's *Test Item Analysis Handout* to examine qualitatively the flagged questions to determine which need revision. The purpose of this paper is to determine if the ACSC Distance Learning Department is doing a good job at test evaluation and whether there are better ways to determine the quality, effectiveness and fairness of multiple-choice questions. This paper compares the *TAD* program to the *ITEMAN* program for quantitative analysis. For qualitative analysis, prior studies and guidelines are compared to those used by ACSC. This paper found that ACSC is doing a good job at test evaluation by using the best threshold values for the Ease Index and Differentiation Index to flag items. This paper recommends that ACSC use the *ITEMAN* software program because of its ease, speed and superior output. This paper recommends that ACSC use the Biserial Correlation Coefficient as well to flag questions. Lastly, this paper recommends that ACSC use Hansen and Dexter's *Item-writing Guidelines* for qualitative review of flagged questions.



## **Chapter 1**

### **Background**

#### **Introduction and Problem Definition**

The multiple-choice question test is perhaps the most popular educational evaluation method used at all levels. The challenge of using this method is designing well-written questions that are reliable and can discriminate the more knowledgeable students from the less knowledgeable students. Every question can be evaluated qualitatively (well written) and quantitatively (reliable and able to discriminate). Qualitative methods can help determine if a question is poorly written. Quantitatively, computer programs provide statistical measures to help determine if a question did not statistically perform well. This paper examines both methods used to evaluate multiple-choice questions.

The Air Command and Staff College Distance Learning Department uses several methods to determine quality, effectiveness, and fairness of multiple-choice questions. Currently, the ACSC Distance Learning Department uses two statistical measures from the *Test Analysis and Development System (TAD)* software to help determine if questions need to be rewritten or deleted. They also use qualitative feedback from students who took the test. The problem of this paper is to determine if the ACSC Distance Learning Department is doing a good job at test evaluation and whether there are better ways to determine the quality, effectiveness and fairness of the multiple-choice questions.

## **Scope of Analysis**

The author selected the *ITEMAN* software program, specifically written to analyze multiple-choice questions, to compare with the *TAD* software program. For quantitative analysis, this paper will compare the statistical measurements used in these software programs. For qualitative analysis, the author will evaluate prior studies on this topic and compare these guidelines to those currently used by ACSC.

## **Thesis**

The ACSC Distance Learning Department currently uses the *TAD* software program's Ease Index and Item Differentiation Index to flag questions for a qualitative review. They consider student opinions on possible unfair questions to also flag questions. The subject matter experts review these flagged questions for quality and fairness. They also use a local guide to determine if questions were properly constructed.

ACSC should consider using the software program *ITEMAN* for quantitative analysis. *ITEMAN* is easy to use, and its output format is superior to the *TAD* program. This output shows the statistical analysis for each question and each answer as well. This output format makes it easy to identify what part of the multiple-choice question may need revision, the correct answer or one of the alternatives.

## **Methodology**

This paper will first provide a review of previous studies completed in the area of multiple-choice question analysis. Next, this paper will compare and analyze the *ITEMAN* program to the *TAD* program. Finally, this paper will evaluate ACSC's current methods and then recommend better methods of multiple-choice question analysis.

## **Chapter 2**

### **Historical Review**

*There is nothing permanent except change.*

—Heraclitus (ca. 540-ca. 480 BC)

#### **Why Use Statistics for Test Item Analysis**

Statistics are very useful in many different applications, and they are particularly useful when analyzing multiple-choice test scores. The assumption in this field of study is that poorly written or “bad” questions’ statistics are fundamentally different to those of “good” questions.<sup>1</sup> After giving a test, an educator can use statistics to identify or “flag” questions with bad statistics in order to examine their quality. Examining a question using statistics is a quantitative method of analysis. Examining a question’s grammar or subject accuracy is a qualitative method of analysis. The ETS Corporation, who analyzes SAT, ACT and CLEP exams for the U.S., only uses quantitative analysis to flag problem questions for review by qualitative methods.<sup>2</sup> ETS will only delete or rewrite a question if it is deficient qualitatively. They will not delete or rewrite a question just because it is quantitatively (statistically) deficient. They are careful to ensure that a question’s statistical analysis is NOT a replacement for sound judgment and knowledge of subject matter.<sup>3</sup> They believe that some good questions can have bad statistics and some bad questions can have good statistics.<sup>4</sup>

## Definitions

This section provides brief descriptions of the paper's commonly used terms, to allow for better understanding when they appear in the remainder of the paper. For complete definitions, please refer to the glossary.

### Ease Index

This is also known as Difficulty Index, Item Difficulty, Percent Correct or “p-value”. It is simply the proportion (or percentage) of students taking the test who answered the item correctly.<sup>5</sup> This value is usually reported as a proportion (rather than percentage), ranging from 0.0 to 1.0. A value of 0.0 would indicate that no one answered the item correctly. A value of 1.0 would indicate that everyone answered the item correctly.

### Differentiation Index

This is also known as Discrimination Index. This statistic is a measure of each test question's ability to differentiate between high scoring and low scoring students. This is computed as: the number of people with highest test scores (top 27%) who answered the item correctly minus the number of people with lowest scores (bottom 27%) who answered the item correctly, divided by the number of people in the largest of the two groups.<sup>6</sup> The higher the number, the more the question is able to discriminate the higher scoring people from the lower scoring people. Possible values range from -1.0 to 1.0. A score of -1.0 indicates that the lowest 27% of the group all answered the question correctly, and the upper 27% of the group all answered the question incorrectly. A score of 1.0 indicates that the upper 27% of the group all answered the question correctly and the lowest 27% of the group answered the question incorrectly.

## Correlation Coefficients

These two coefficients are also known as Discrimination Coefficients.

**Biserial Correlation Coefficient.** This statistic correlates overall test scores to the correct answering of an individual test item (question).<sup>7</sup> In other words, it is a measurement of how getting a particular question correct correlates to a high score (or passing grade) on the test. Possible values range from  $-1.0$  to  $1.0$ . A score of  $-1.0$  would indicate that all those who answered the question correctly scored poorly on (or failed) the test. A score of  $1.0$  would indicate that those who answered the question correctly scored well on (or passed) the test.

**Point Biserial Correlation Coefficient.** This statistic is a measure of the capacity of a test item (question) to discriminate between high and low scores.<sup>8</sup> In other words, it is how much predictive power an item has on overall test performance. Possible values range from  $-1.0$  to  $1.0$  (the maximum value can never reach  $1.0$ , and the minimum can never reach  $-1.0$ ). A value of  $0.6$  would indicate the question has a good predictive power, i.e., those who answered the item correctly received a higher average grade compared to those who answered the item incorrectly. A value of  $-0.6$  would indicate the question has a poor predictive power, i.e., those who answered the item incorrectly received a higher average grade compared to those who answered the item correctly.

## Quantitative Factors

Listed above, these four statistical measurements are the most widely used quantitative methods to evaluate multiple-choice questions. The *TAD* software also offers another statistical measurement, the Brennan's B Coefficient. This statistic is used when the instructor identifies a mastery criterion group within the group of students being tested.<sup>9</sup> The "Masters" (upper scoring) group is usually the group that passes the test; whereas the "Nonmasters" (lower

scoring) group is usually the group that fails the test.<sup>10</sup> This statistic is calculated by taking the number of Master students answering the item correctly, subtracting the number of Nonmaster students answering the item correctly and then dividing by the total number of students.<sup>11</sup> Since such a high percentage of the ACSC students pass the tests, this statistic is not a very useful quantitative measure. Therefore, the Brennan's B Coefficient statistic will not be considered in this paper.

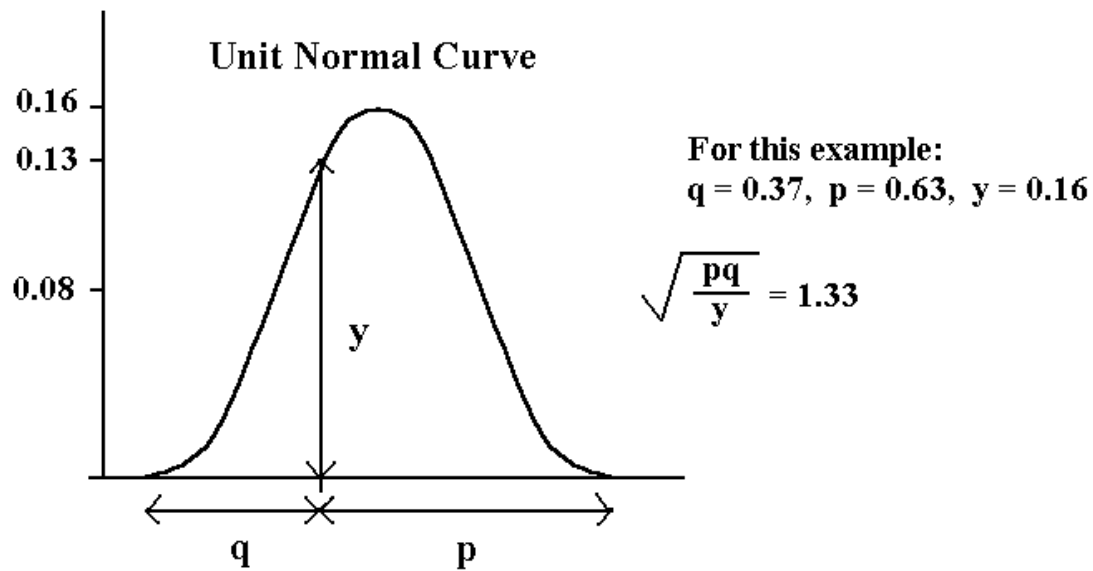
The **Ease Index** is an excellent statistic to measure a question's difficulty. What value would be an ideal Ease Index value? Using a four-alternative multiple-choice test, Thompson and Levitow in 1985 reported that the ideal difficulty of an item would be halfway between the percentage of pure guess (25%) and everyone answering correctly (100%)<sup>12</sup>. This equals a percentage of 62.5%, or an Ease Index of 0.625. For a three-alternative multiple-choice test, the ideal Ease Index value would be  $(33\%+100\%)/2$ , equaling 66.5% 0.665. For a five-alternative multiple-choice test, the ideal Ease Index value would be  $(20\%+100\%)/2$ , equaling 60% or 0.60.

The strength of the **Differentiation Index** (DI) is that most people can understand it more easily than the correlation coefficients. The formula for this index is simple to compute and simple to understand. A few authors point out that this statistic does have a drawback. Susan Matlock-Hetzel states that the advantage of the correlation coefficients over the DI is that every person taking the test is used to compute the coefficients and only 54% are used to compute the DI.<sup>13</sup> Dr. Thomas Renckly, creator of *TAD*, states that "the DI only makes use of two-thirds of the available test scores, as the DI makes no use of middle-third of scores."<sup>14</sup> His program calculates the DI using the upper third and lower third of the group compared to the more commonly used top 27% and bottom 27%.<sup>15</sup> Even one author, suggested using 25% instead of

27% in order to maximize the discrimination of the DI.<sup>16</sup> In that case, 50% of the scores would not be used.

Dr. Thomas Renckly states that the **Point Biserial Correlation Coefficient** (Rpbis) is a more stable alternative to the Differentiation Index.<sup>17</sup> He argues that the higher stability derives from “(1) this coefficient makes use of all test data and (2) the computation does not depend on arbitrary cutoff values (as does the DI)”.<sup>18</sup> One can also apply this statistic to each alternative of the question to help determine poor distractors (incorrect answers). However, Attali and Fraenkel show that while it is sound to use the Rpbis calculation for the correct answer, one should not use it for the distractors.<sup>19</sup> They argue that the Rpbis for the distractor is unreliable. This is because of those who did not select the distractor, there is no distinction between those who chose the correct answer and those who did not choose the correct answer.<sup>20</sup> They suggest that distractors not be eliminated solely based on the Rpbis value for that distractor.

The **Biserial Correlation Coefficient** (BIS) is usually preferred over the Point Biserial as a criterion measure for a multiple-choice question because the Point Biserial depends heavily on question difficulty.<sup>21</sup> Attali and Fraenkel say that the Biserial is usually preferred as a criterion measure for the correct alternative “because its values are not sensitive to item difficulty”.<sup>22</sup> The relationship between the Biserial (BIS) and Point Biserial (Rpbis) is expressed by the equation:  $BIS = Rpbis \times (\sqrt{pq}/y)$ , where  $p$  is the number of those answering the item correctly,  $q$  is the number of those answering incorrectly, and  $y$  is the height of the ordinate of the unit normal curve at the point of division between the  $p$  and  $q$  proportions of cases.<sup>23</sup> See figure below for a graphic depiction of  $p, q, y$  and the unit normal curve.



**Figure 1 Unit Normal Curve and Values of  $q$ ,  $p$  and  $y$**

The value of  $(\sqrt{pq/y})$  can vary from 1.25 to 3.73.<sup>24</sup> Therefore, the value of the Biserial is always greater than the Point Biserial; and this difference increases as the difference between  $p$  and  $q$  increases. The main point here is that the Biserial Coefficient is not dependent on question difficulty.

### **Qualitative Factors**

One can say that the quantitative factors are the “math” methods and that the qualitative factors are the “English” methods of evaluating multiple-choice questions. The number of answers, the sequencing of the questions, proper question construction and student feedback are the qualitative factors addressed in this paper.

The first factor a test question writer must consider is how many answers each question should have. This has been a subject of much research for more than sixty years. The objective is to have a correct answer that most students select, along with attractive alternatives that some



students select. Any alternative that is selected by very few is called a non-functioning alternative and should be rewritten or deleted.

Gregory Cizek and Denis O'Day researched the effects of removing a non-functioning alternative from a group of five-option test questions. They defined a non-functioning option as one selected by 4% or less of the examinee group. They found that questions with a non-functioning option removed were equally reliable as the same question in the original five-option format.<sup>25</sup> In 1993, Thomas Haladyna and Steven Downing also examined the question of how many options are enough. Their paper examined both theoretical and empirical evidence over a period of more than forty years, analyzing data on two, three, four and five option test questions. Their research supported the use of three-option test questions as optimum.<sup>26</sup>

With the questions written, the test writer then must decide in what order the questions should appear on the test. The test writer can choose to order the questions in a forwardly ordered way (i.e., corresponding with sequence of class coverage) or in a random order. Marshall Geiger and Kathleen Simons examined both the test scores and time needed to complete for both forwardly and randomly ordered tests. They found that there were no differences in scores or completion times between forwardly ordered and randomly ordered tests of the same test questions.<sup>27</sup> Therefore, a teacher can use a number of versions of the same exam knowing that these different versions will not affect students' scores.

Proper question construction is an important factor in determining the quality and effectiveness of a question. Question writing guidelines help the writer construct a question correctly the first time. One can also use these guidelines to help determine how to change a "bad" question around to make it a "good" question. There are many articles and books on multiple-choice question writing that have been published in the last fifteen years alone.<sup>28</sup> James

Hansen and Lee Dexter reviewed many of these publications in their research and compiled the seventeen most-cited guidelines.<sup>29</sup> ACSC uses the Maxwell Academic Instructor School *Test Item Analysis Handout* guidelines to construct a proper question or assess if a question is properly constructed. These two collections of writing guidelines will be compared and analyzed in Chapter 3.

### **ACSC Distance Learning Department's Current Methods**

The ACSC Distance Learning Department uses two quantitative methods and two qualitative methods to analyze questions for their effectiveness. The department uses three of these four methods to flag questions for a qualitative review by subject matter experts. The department also uses the Academic Instructor School handout question construction guidelines to help improve a question's quality and effectiveness.

The ACSC Distance Learning Department uses two quantitative methods, the Ease Index and the Differentiation Index, to flag questions for review. Their threshold for the Ease Index is 50% (or 0.5). If a question scores an EI value lower than 50%, then the subject matter expert will review the question for difficulty. Similarly, their threshold for the Differentiation Index is 0% (or 0.0). Questions scoring a DI value lower than 0.0 are reviewed by the subject matter experts to determine if they can be rewritten to be more effective. The department prefers that questions have a Differentiation Index score of 0.2 or better.

Along with these two statistical measures, the Distance Learning Department uses feedback from the students to help determine questions that seemed unfair. After taking the examination, students can indicate which questions they felt were "bogus". This feedback is then evaluated to flag additional questions for qualitative review by the subject matter experts.

The ACSC Course Directors write the multiple-choice questions used for all the distance learning tests. They receive a class by Maxwell's Academic Instructor School; where they learn about both qualitative analysis and quantitative analysis methods. The school's handout, *Test Item Analysis Handout*, contains eleven test question construction guidelines on ways to improve questions and their effectiveness.<sup>30</sup> Course Directors use these guidelines to quality check new questions as well.

### Notes

<sup>1</sup> Kehoe, Jerard. (1995). "Basic Item Analysis for Multiple-Choice Tests," <http://ericae.net/digests/tm9511.htm>, p. 1.

<sup>2</sup> Zieky, Michael. Email concerning ETS's item analysis, 06 Nov., 2000. Email address: [mzieky@ets.org](mailto:mzieky@ets.org), p. 1.

<sup>3</sup> Ibid, p. 1.

<sup>4</sup> Ibid, p. 1.

<sup>5</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>. p.1.

<sup>6</sup> Ibid, p.6

<sup>7</sup> Ferguson, George A. (1976). *Statistical Analysis in Psychology and Education* (4<sup>th</sup> ed.), New York, NY, McGraw-Hill, p. 418.

<sup>8</sup> Ibid, p. 416.

<sup>9</sup> Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

<sup>10</sup> Ibid.

<sup>11</sup> Ibid.

<sup>12</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>, p.4

<sup>13</sup> Ibid, pp. 7-8.

<sup>14</sup> Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

<sup>15</sup> Ibid

<sup>16</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>, p.6

<sup>17</sup> Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

<sup>18</sup> Ibid.

<sup>19</sup> Attali, Yigal and Fraenkel, Tamar. "The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative," *Journal of Educational Measurement*, vol. 37, no. 1, (Spring 2000), p. 77.

<sup>20</sup> Ibid, p. 78.

## Notes

<sup>21</sup> Zieky, Michael. Email concerning ETS's item analysis, 06 Nov., 2000. Email address: mzieky@ets.org .

<sup>22</sup> Attali, Yigal and Fraenkel, Tamar. "The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative," *Journal of Educational Measurement*, vol. 37, no. 1, (Spring 2000), p. 80.

<sup>23</sup> Ferguson, George A. (1976). *Statistical Analysis in Psychology and Education* (4<sup>th</sup> ed.), New York, NY, McGraw-Hill, p. 418.

<sup>24</sup> Ibid, p. 419.

<sup>25</sup> Cizek, Gregory J., O'Day, Denis M. "Further Investigation of Nonfunctioning Options in Multiple-Choice Test Items," *Educational & Psychological Measurement*, vol. 54, issue 4, (Winter 1994), p. 867.

<sup>26</sup> Haladyna, Thomas M. and Downing Steven M. "How Many Options is Enough for a Multiple-Choice Test Item?" *Educational & Psychological Measurement*, vol. 53, issue 4, (Winter 1993),p. 1004.

<sup>27</sup> Geiger, Marshall A. and Simmons, Kathleen A. "Intertopical Sequencing of Multiple-Choice Questions: Effect on Exam Performance and Testing Time" *Journal of Education & Business*, vol. 70, issue 2, (Nov/Dec 1995), p.90.

<sup>28</sup> Hansen, James D. and Dexter, Lee. "Quality Multiple-Choice Test Questions: Item-Writing Guidelines and an Analysis of Auditing Testbanks," *Journal for Education for Business*, vol. 73, no. 2, (Nov 1997), p. 94, Heldref Publications.

<sup>29</sup> Ibid.

<sup>30</sup> *Test Item Analysis Handout*, AI-641-c, 02-00, Academic Instructor School, Maxwell AFB, AL.

## Chapter 3

### Analysis

*Those who cast their vote decide nothing; those who count the votes decide everything.*

—Joseph Stalin

#### ***TAD* Software Program**

*TAD* version 5.49.35 is an excellent program for all aspects of scoring, analyzing and grading tests. This program can do much more than just statistical analysis of multiple-choice questions. For this paper, the author will only analyze those portions of the program that deal specifically with the statistical measures defined in Chapter 2.

#### **Inputs and Controls**

*TAD* allows the user to input either manually or automatically. To input data manually, the user enters the correct answers first, then enters each student and only the answers that they missed. Entering only the missed answers speeds up the input process; however, manual inputting the data is time consuming. Since ACSC receives the test data in a computer file already, they would use this manual method only as a backup method if the data file was lost or corrupted.

To enter data automatically, the user first identifies the file for analysis. The program brings up one string of data; the user then identifies the student name and/or number and where the

answers begin and end. The program then brings up a menu of different options for the user to select from before it translates the file into a format necessary for analysis.

The controls in the *TAD* program are simple and easy to understand. The menus are the standard Window's pull-down menus. The help file is comprehensive and easy to use. In addition to the standard help files, *TAD* also offers help cards and automated tutorials. These additional features help to clarify certain program features. The program also allows the user to set threshold values for the item statistics: Ease Index, DI and Point Biserial. This makes it easy for the user to find questions that fall below set thresholds. One important feature this program does not offer is the calculation of the Biserial Correlation Coefficient. This statistic is preferred over the Point Biserial Correlation Coefficient since the Point Biserial depends heavily on item difficulty.<sup>1</sup>

## Output Format

The *TAD* program displays its output in a number of different windows. The window that is of most importance for question analysis is the "Item Statistics" window. The figure below is a sample of the data presented in this window.

TestBank ID Code	Ease Index	Diff. Index	Item Rpbis	Alt. A Rpbis	Alt. B Rpbis	Alt. C Rpbis
1	93	0.04	0.11	-0.21	-0.02	0.11
2	74	0.41	0.42	-0.37	0.42	-0.16
3	40	0.33	0.26	-0.25	-0.05	0.26
4	100	0	0.00	0.00	0.00	0.00
5	92	0.12	0.24	-0.23	-0.05	0.24

**Figure 2 *TAD* "Item Statistics" window sample output<sup>2</sup>**

This output's format is simple and allows the user to see all three statistics at the same time. Additionally, the Point Biserial Correlation Coefficients for each answer are included. It is

worth noting here that when all students answer the question correctly, (see **Testbank ID Code #4** in Figure 1 above) the DI and Point Biserials values are zero.

### **User Friendliness**

The *TAD* software program is very user friendly. The help files, automated tutorials, help cards, *TAD* Tips and even a *TAD* web site allow the user to find the answer they need quickly. Additionally, the creator of the program, Dr. Thomas Renckly, works at Maxwell AFB, and is available to answer questions as well.

### ***ITEMAN* Software Program**

The *ITEMAN* software program analyzes multiple-choice questions and can also compute and record test scores. The program is advertised as a “Classical Item Analysis” program.<sup>3</sup> The version number is 3.50, available on the Internet at [www.assess.com](http://www.assess.com).<sup>4</sup> This program offers four statistical measures: Ease Index (called “Proportion Correct”), Discrimination Index, Biserial and Point Biserial Correlation Coefficients. The *TAD* program offers three of these statistics, but does not offer the Biserial Coefficient.

### **Inputs and Controls**

*ITEMAN* requires that the input data file be formatted in ASCII (text-only) files. Most data files produced by optical scanning devices are very close to the format that *ITEMAN* requires, with the exception of the four lines that must be added at the beginning.<sup>5</sup> These lines contain the control line, the key, number of alternatives, etc. The program can process up to a 750-item test with unlimited number of students.

The user can also manually generate a data file using the edit menu in *ITEMAN*, which is similar to Windows’ Notepad program. Unlike *TAD*, this manual method would require the user

to type in every answer - very time consuming! Again, ACSC would only use this manual method as a backup method if the file of data were lost.

*ITEMAN's* controls are few in number and very simple to use. The program offers five pull down menus and five buttons. The user first selects the configure menu or button to identify the file and select the options desired for analysis. The user then selects the analyze menu or button. The user can view or print the output file by clicking on the view button or print button. These buttons appear after the analysis is complete.

### Output Format

*ITEMAN* produces an output file, score file (if desired) and statistics file (if desired). The output file contains the statistical measures, and displays them not only for each question, but for each alternative as well. Figure 2 is a sample from the output file.

Seq. No.	Scale Item	Item Statistics			Alternative Statistics					Key
		Prop. Correct	Disc. Index	Biser.	Alt.	Prop. Total	Endorsing Low	High	Biser.	
1	0-1	.79	.37	.53	1	.16	.29	.04	-.36	
					2	.04	.10	.00	-.52	
					3	.01	.03	.00	-.87	
					4	.79	.59	.96	.53	*
					Other	.00	.00	.00		
2	0-2	.69	.69	.79	1	.04	.10	.00	-.61	
					2	.69	.29	.97	.79	*
					3	.05	.15	.00	-.63	
					4	.22	.46	.03	-.54	
					Other	.00	.00	.00		
3	0-3	.94	.18	1.00	1	.04	.12	.00	-.93	
					2	.01	.03	.00	-.74	
					3	.94	.82	1.00	1.00	*
					4	.01	.04	.00	-.79	
					Other	.00	.00	.00		

**Figure 3 *ITEMAN* Sample Output Using EI, DI and Biserial Coefficients<sup>6</sup>**



This output lists the proportions of 1) the total number of students selecting, 2) the bottom 27% of the group selecting, and 3) the top 27% of the group selecting for each alternative.<sup>7</sup> The output also lists the Biserial Coefficients for each alternative. The asterisk denotes which alternative is the correct answer. This format allows the user to examine alternatives by comparing the high scoring students selecting versus the low scoring students selecting. This easily allows the user to identify alternatives that are attracting too many high scoring students, indicating the alternative may need revision.

### **User Friendliness**

*ITEMAN* is a very user-friendly program partly because it has so few controls and only one output file. The DOS version of the program (which comes with the purchase of the Windows version) can be set up to run in batch mode at night, so the computer is available during the day. A user would use the batch mode if they had large datasets to analyze. The sample file, included with the trial program, contains 400 students taking a 20-question test. *ITEMAN* analyzed this file on a Pentium 200 MHz computer in approximately 1 second. Most users would not need to use the batch mode.

### **Quantitative Analysis**

The ideal **Ease Index** value for a question depends on how many alternatives there are. For a 3-option question 0.665 is ideal, and for a 4-option question 0.625 is ideal.<sup>8</sup> No source used for this paper examined or recommended what a good EI threshold value should be to flag problem questions. Like the ideal value, the threshold value depends on how many alternatives. Mathematically, the best EI threshold value would be halfway between the ideal Ease Index value and the “pure guess” Ease Index value.

The **Differentiation Index** is a popular statistical measure because of the ease of computation compared to the correlation coefficients. Though the DI does not use all of the test scores in its calculation, it is a good measure to see if the question is discriminating between the highest scoring students and the lowest scoring students. Christina Ballantyne states that a DI of +0.30 indicates that an item is working well, but a DI of +0.20 is not bad.<sup>9</sup> Susan Matlock-Hetzel supports that a DI of +0.20 to +0.29 indicates the item needs some revision, and that a DI of +0.19 or less indicates the item needs major revision or should be eliminated.<sup>10</sup> Thus, a threshold value to flag questions for revision should probably have a DI value lower than +0.20.

The **Point Biserial Correlation Coefficient** is a controversial statistical measure simply because it depends heavily on item difficulty. Jerard Kehoe states that items in which 85% of the examinees answered correctly have a reduced power to discriminate.<sup>11</sup> Depending on the question, the Point Biserial value may not be useful. Proponents of the Point Biserial, like Susan Matlock-Hetzel and Dr. Thomas Renckly, argue for its consideration because the statistic uses all student scores in its calculation. Jerard Kehoe is the only author to suggest a threshold value. He suggests that a question having a Point Biserial value of +0.15 or less should probably be rewritten.<sup>12</sup>

The **Biserial Correlation Coefficient** is usually preferred over the Point Biserial because its values are not sensitive to item difficulty.<sup>13</sup> Using the *ITEMAN* program to analyze the same data using the Biserial and Point Biserial, it is easy to see the Biserial does not fluctuate in value as greatly as does the Point Biserial (See Appendix A). Michael Zieky, of the ETS Corporation, says that they use a Biserial threshold value of +0.30 and lower to flag questions for qualitative review.<sup>14</sup>

Along with running statistics for each question, Susan Matlock-Hetzel references Millman and Greene's research and recommends the use of statistics like the Differentiation Index or correlation coefficients for each alternative as well.<sup>15</sup> She states that distractors should have a correlation coefficient less than the correct answer and preferably negative.<sup>16</sup> She suggests the subject matter expert should carefully examine distractors with large positive values.<sup>17</sup>

### **Qualitative Analysis**

Previous research on the most suitable number of alternatives for a multiple-choice question supports both a 4-option question and a 3-option question. These studies strongly support the idea that any non-functioning distractor should be eliminated. Additionally, Knowles and Welch support the use of "none-of-the-above" (NOTA) as a valid alternative. Their research shows that the use of the NOTA alternative would only decrease the students' scores by 1.00 on a 100-question test.<sup>18</sup> Therefore, an instructor could change a 3-option test to a 4-option test simply by adding the NOTA alternative to each question. The main point here is that each question should have at least three quality alternatives, with portions of the students selecting each of the two distractors. Ideally, a greater number of the less knowledgeable students should select the distractors compared to the more knowledgeable students.

Concerning question order, an instructor need not worry about the order of questions on the test. As discussed earlier, Marshall Geiger and Kathleen Simons found that there were no differences in scores or completion times between forwardly ordered and randomly ordered tests of the same test questions.<sup>19</sup> Thus, an instructor can use several versions of a test from a question bank, and ensure fairness regardless of the version.

Writing guidelines are important to ensure instructors properly construct questions, both grammatically and structurally. Hansen and Dexter examined many studies on question writing

guidelines and compiled a comprehensive list of rules and guidelines in their 1997 article. Many of these guidelines are the same as the one listed in the *Test Item Analysis Handout* of the Maxwell Academic Instructor School. Both lists are located in Appendix B. Maxwell's guidelines are brief, with no examples. Hansen and Dexter's list is detailed and comprehensive, containing nine more guidelines. These items are: 8. (a) through (f) and 10 through 12 (see Appendix B). The ACSC Distance Learning Department should consider these items for their review of questions.

### **ACSC Distance Learning Department's Current Methods Analysis**

The purpose of this paper is to answer two questions. The first is to determine if the ACSC Distance Learning Department is doing a good job at test evaluation. This section will evaluate their current methods.

Currently, ACSC uses two statistical measures, the Ease Index and Differentiation Index. They use the *TAD* software program to analyze the student test data and calculate these two statistical measures.

ACSC uses an Ease Index threshold value of 0.50 to flag questions for review. Since their questions have three alternatives, the ideal Ease Index is 0.665. The Ease Index for merely guessing is 0.333. The best threshold value would be halfway between the ideal Ease Index and the "pure guess" value. This value is  $(0.665 + 0.333)/2$  or 0.499, which rounds to 0.50. Therefore, ACSC is using the best Ease Index threshold value of 0.50. If they were to use a 4-option test, the best Ease Index threshold value would be  $(0.625 + 0.25)/2$ , equaling a value of 0.4375 or 0.44 when rounded to two digits.

ACSC uses a Differentiation Index threshold value of 0.0 to flag questions for review. Christina Ballantyne states a DI of +0.30 indicates that an item is working well, but a DI of

+0.20 is not bad.<sup>20</sup> Susan Matlock-Hetzel states a DI of +0.20 to +0.29 indicates the item needs some revision, and a DI of +0.19 or less indicates the item needs major revision or should be eliminated.<sup>21</sup> One author is saying +0.20 is not bad; and the other author says +0.20 indicates a need for revision. This difference in opinion is most likely due to the difference in student populations and scores analyzed in each study.

A critical aspect of the Differentiation Index is that it also relies on item difficulty. If a group of students score relatively high on many questions, then the DI values of those questions will be very small, often falling between the values of  $-0.05$  and  $+0.20$ . The ACSC student population is such a group, since their Ease Indexes are usually high, typically around 0.80 (80%) or higher (See *TAD* output in Appendix A).

Analyzing the *TAD* output in Appendix A, one can see that 36 of the 62 questions have an Ease Index of 0.80 (80%) or higher. The number of questions with DI values between  $-0.05$  and  $+0.20$  is 37, almost 60% of the total number of questions! Using a DI threshold value of  $+0.20$  on this test data would not be of much help, since it would flag 39 of the 62 questions (63% of the questions). Using a DI threshold value of 0.0, only six of the 62 questions fall below this threshold (about 10%), which is a reasonable amount.

By definition, a DI value of 0.0 means the same number of lower scoring students and higher scoring students answered the question correctly. A negative DI value means more lower scoring than higher scoring students are answering the question correctly. Considering both the data and the DI definition, a DI threshold value of 0.0 is the logical value. Thus, ACSC is using the best DI threshold value for their student population.

The ACSC Distance Learning Department uses student feedback as a qualitative method to flag questions for review by subject matter experts. This method is a sound way to flag problem

questions for review. Students usually have a good idea when a question seems unfair or is improperly written. Another advantage of this method is that it does not require any statistical analysis; feedback is quick and direct. ACSC should continue to use student feedback to help flag questions for review.

ACSC also uses the Maxwell Academic Instructor School's *Test Item Analysis Handout* to help write questions and review flagged questions for structural problems and content. However, the section in the handout dealing with test construction is quite brief. Since Hansen and Dexter's *Item-writing Guidelines* are more detailed and comprehensive, ACSC should include these guidelines for qualitative review.

The ACSC Distance Learning Department uses three versions of each test with 3-option questions. As discussed previously, research supports either 3-option questions or 4-option questions as optimum. Geiger and Simons' research showed that there is no difference in test scores or completion times between forwardly ordered and randomly ordered tests. ACSC's use of three versions of each test with 3-option questions is very sound, as the research supports this.

### Notes

<sup>1</sup> Zieky, Michael. Email concerning ETS's item analysis, 06 Nov. 2000. Email address: mzieky@ets.org .

<sup>2</sup> Output used with permission, Thomas R. Renckly, Jan. 12, 2001. Output from *Test Analysis & Development System (TAD) version 5.49*. CD-ROM.(1990-2000).

<sup>3</sup> *ITEMAN* Online Manual. Assessment Systems Corporation. <http://www.assess.com> .

<sup>4</sup> Ibid.

<sup>5</sup> Ibid.

<sup>6</sup> Output used with permission, David J. Weiss, President, Assessment Systems Corporation, Jan. 11, 2001. Output from *ITEMAN* software program, demonstration version 3.50 Assessment Systems Corporation. (1995). <http://www.assess.com> .

<sup>7</sup> *ITEMAN* Online Manual. Assessment Systems Corporation. <http://www.assess.com> .

<sup>8</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>. p. 4.

<sup>9</sup> Ballantyne, Christina. (2000). "Multiple-Choice Tests: Test Scoring and Analysis" <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html> p. 5.

## Notes

<sup>10</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>. p. 7.

<sup>11</sup> Kehoe, Jerard. (1995). "Basic Item Analysis for Multiple-Choice Tests," <http://ericae.net/digests/tm9511.htm> p. 1.

<sup>12</sup> Ibid, p. 2.

<sup>13</sup> Attali, Yigal and Fraenkel, Tamar. "The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative," *Journal of Educational Measurement*, vol. 37, no. 1, (Spring 2000), p. 80.

<sup>14</sup> Zieky, Michael. Email concerning ETS's item analysis, 06 Nov. 2000. Email address: [mzieky@ets.org](mailto:mzieky@ets.org).

<sup>15</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>. p. 8.

<sup>16</sup> Ibid.

<sup>17</sup> Ibid.

<sup>18</sup> Knowles, Susan L. and Welch, Cynthia A. "A Meta-Analytic Review of Item Discrimination and Difficulty in Multiple-Choice Items Using "None-of-the-Above"," *Educational & Psychological Measurement*, vol. 52, issue 3, (Fall 1992), p.574

<sup>19</sup> Geiger, Marshall A. and Simmons, Kathleen A. "Intertopical Sequencing of Multiple-Choice Questions: Effect on Exam Performance and Testing Time" *Journal of Education & Business*, vol. 70, issue 2, (Nov/Dec 1995), p.90.

<sup>20</sup> Ballantyne, Christina. (2000). "Multiple-Choice Tests: Test Scoring and Analysis" <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html> p. 5.

<sup>21</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html> p. 7.

## **Chapter 4**

### **Conclusions**

#### **Summary of Findings**

In summary, quantitative methods using statistical analysis of data from multiple-choice questions are widely used and popular. The most commonly used statistical measures are the Ease Index, Differentiation Index, Biserial Correlation Coefficient, and Point Biserial Correlation Coefficient. Educators use these statistical measures to flag questions that are not “statistically” performing well. These flagged questions are given to subject matter experts for review using qualitative methods. Most educators do not simply throw out a question for having of bad statistics.

Subject matter experts use qualitative methods to analyze the structure of the sentence and content for accuracy. They use question writing guidelines as a checklist to correctly write and rewrite test questions. Student feedback is another method widely used by educators to flag questions that students thought were unfair or poorly written.

#### **ACSC Distance Learning Department’s Current Methods**

ACSC’s Distance Learning Department uses two quantitative methods and two qualitative methods to analyze questions for their effectiveness. They use the *Test and Analysis System (TAD)* software with an Ease Index threshold value of 0.50 and a Differentiation Index threshold



value of 0.0 to flag questions for review. This paper found that the best Ease Index threshold value was, in fact, 0.50. Since ACSC students' test scores are higher than normal, using a DI threshold value of 0.0 proved to be the best threshold value for ACSC student populations.

ACSC's qualitative methods include using student feedback to flag questions for review and using the Maxwell Academic Instructor School's *Test Item Analysis Handout* to help write and review flagged questions for structural problems and content. This paper supports the use of student feedback to flag questions and the question writing guide to help correctly write, review and rewrite questions.

### ***TAD* versus *ITEMAN***

The ACSC Distance Learning Department's statistical analysis considers two statistical measures, the Ease Index and Differentiation Index. Because of these limited needs, ACSC should use the *ITEMAN* software program. *ITEMAN* is easier and faster to use compared to the *TAD* software program.

*ITEMAN's* output is superior as well. This output provides statistics for each question, and statistics for each alternative. The analysis of alternatives includes the proportions of students of the top 27% and bottom 27% that selected each alternative, which the *TAD* program does not. This data is valuable, as it will tell the subject matter expert which alternatives are working well and which need revision.

Although not currently used by ACSC, *ITEMAN* also offers the calculation of Biserial and Point Biserial Correlation Coefficients. *TAD* only offers the Point Biserial. Previous research analyzed in this paper recommends the use of the Biserial over the Point Biserial because the Point Biserial depends heavily on question difficulty.

## **Recommendations**

### ***ITEMAN* as the Preferred Program**

This paper recommends the *ITEMAN* software program as the preferred software program for use by the ACSC Distance Learning Department. The *ITEMAN* program is specifically designed for item analysis only. *ITEMAN* is easier and faster to use compared to the *TAD* program. If desired, the user can configure the program to work by batch mode at night, freeing up computer time during work hours. *ITEMAN's* proportions of the top 27% and bottom 27% selecting each alternative show which alternatives are working well and which are not. *ITEMAN* also offers the Biserial Coefficient, the preferred correlation coefficient, which *TAD* does not offer.

If the ACSC Distance Learning Department decides not to use the *ITEMAN* program, then they should continue using the *TAD* program. In this case, ACSC should ask Dr. Thomas Renckly if it is possible for him to include question alternative analysis similar to what *ITEMAN* provides. They should also ask him to add the Biserial Correlation Coefficient analysis to the *TAD* program as well.

### **Quantitative Measurements**

ACSC's Distance Learning Department is doing a good job by using the Ease Index and Differentiation Index to flag questions for review. They should continue using the threshold value of +0.50 for the Ease Index and the threshold value of 0.0 for the DI.

The Differentiation Index does have two drawbacks. Depending on which software program you use, the DI calculation may leave out 33% or 46% of the students' scores in its calculation. The DI, like the Point Biserial, relies on item difficulty. For questions with high Ease Index values, the DI values may be lower than 0.0.

Because of the possibility of erroneously low DI values for certain questions, this paper recommends that ACSC use the Biserial Correlation Coefficient in addition to the EI and DI for quantitative analysis. The Biserial Coefficient is a more stable measurement, as it does not vary as greatly as does the Point Biserial and Differentiation Index. Additionally, the Biserial does use all of the students' scores for its calculation. Initially, ACSC should use a Biserial threshold value of +0.30. This value is what the ETS Corporation uses; and no other source recommended another value. Of course, ACSC can adjust the threshold if their experience proves that a different value is better.

### **Qualitative Guidelines**

The use of student feedback by ACSC to flag questions is a sound practice and should continue this practice. Student feedback is easy to collect and usually a good indicator of which questions have problems and need revision. Since feedback is used only as a flag and not as a basis to throw out questions, ACSC should continue to use it.

This paper recommends that ACSC use Hansen and Dexter's *Item-writing Guidelines* instead of the Maxwell *Test Item Analysis Handout* they currently use. Hansen and Dexter's guidelines are more detailed and comprehensive compared to Maxwell's handout.

ACSC should continue using the three versions of 3-option question exams for each course. Research supports both 3-option and 4-option exams as optimum, so they could expand their questions to 4-option if they desired. Research also supports the use of the "none-of-the above" (NOTA) as a viable alternative, which could easily turn a 3-option exam into a 4-option exam.

## Appendix A

### Sample Outputs From Software Programs

#### *TAD* Sample Output Using EI, DI, and Point Biserial Correlations

The sample output below is from the *TAD* software program using a 2-year-old data file of ACSC students' test scores. The Ease Index, Differentiation Index, and Point Biserial Correlation Coefficient values are displayed. Items with a flag indicate the item has a DI value less than 0.0 or an EI value less than +0.50.

Flagged Item	TestBank ID Code	Ease Index	Diff. Index	Item Rpbis	Alt. A Rpbis	Alt. B Rpbis	Alt. C Rpbis
	1	93.00	0.04	0.11	-0.21	-0.02	0.11
	2	74.00	0.41	0.42	-0.37	0.42	-0.16
►	3	40.00	0.33	0.26	-0.25	-0.05	0.26
	4	100.00	0.00	0.00	0.00	0.00	
	5	92.00	0.12	0.24	-0.23	-0.05	0.24
	6	99.00	0.04	0.05	-0.05	0.05	
	7	85.00	0.12	0.12	0.00	0.12	-0.05
	8	75.00	0.16	0.20	0.20	-0.30	-0.05
	9	93.00	0.00	-0.01	0.10	0.00	-0.01
	10	77.00	0.16	0.19	-0.02	0.19	-0.19
	11	82.00	0.16	0.19	0.19	0.00	-0.19
	12	97.00	0.08	0.37	-0.37	0.37	
►	13	93.00	-0.12	-0.18	0.05	0.17	-0.18
	14	90.00	0.08	0.10	0.00	0.10	-0.11
	15	82.00	0.04	0.13	0.13	-0.09	-0.09
	16	62.00	0.25	0.14	0.14	-0.06	-0.12
	17	95.00	0.00	0.10	0.10	-0.10	
	18	92.00	0.25	0.38	-0.30	-0.23	0.38
►	19	48.00	-0.08	0.04	0.16	0.04	-0.23
	20	73.00	0.53	0.48	-0.31	-0.32	0.48

**Figure 4** *TAD* Sample Output using EI, DI and Point Biserial (Part 1 of 2)<sup>1</sup>

Flagged Item	TestBank ID Code	Ease Index	Diff. Index	Item Rpbis	Alt. A Rpbis	Alt. B Rpbis	Alt. C Rpbis
	21	88.00	0.04	0.20	-0.20	0.20	
	22	84.00	0.16	0.12	-0.11	-0.05	0.12
	23	74.00	0.29	0.26	-0.11	0.26	-0.31
	24	96.00	0.08	0.08	0.08	-0.08	
	25	100.00	0.00	0.00	0.00		
	26	66.00	0.25	0.23	-0.06	0.23	-0.21
	27	90.00	0.16	0.30	-0.26	0.30	-0.16
►	28	32.00	0.00	-0.01	-0.01	0.13	-0.09
	29	93.00	0.12	0.36	-0.35	0.36	-0.17
	30	88.00	0.12	0.16	0.16	0.00	-0.16
	31	75.00	0.25	0.34	-0.27	-0.18	0.34
►	32	81.00	-0.04	-0.02	-0.04	-0.02	0.12
	33	73.00	0.21	0.10	0.05	-0.15	0.10
►	34	44.00	0.21	0.18	-0.13	-0.11	0.18
	35	81.00	0.25	0.35	-0.31	-0.26	0.35
	36	96.00	0.00	-0.04	-0.04	0.00	0.04
	37	68.00	0.33	0.23	-0.26	0.05	0.23
	38	92.00	0.16	0.14	-0.18	0.14	-0.06
	39	67.00	0.12	0.18	-0.16	0.18	-0.06
	40	82.00	0.29	0.37	0.37	-0.34	-0.13
►	41	99.00	-0.04	-0.08	0.08	-0.08	
	42	53.00	0.29	0.33	-0.30	0.33	-0.08
	43	71.00	0.21	0.27	-0.17	-0.18	0.27
	44	93.00	0.16	0.36	-0.29	-0.26	0.36
	45	79.00	0.12	0.17	-0.21	0.17	-0.09
	46	70.00	0.12	0.04	0.04	0.02	-0.11
	47	62.00	0.41	0.28	-0.21	-0.29	0.28
	48	86.00	0.25	0.32	-0.22	-0.23	0.32
	49	84.00	0.29	0.44	-0.06	0.44	-0.45
	50	96.00	0.08	0.28	-0.29	0.28	-0.13
	51	79.00	0.25	0.22	-0.28	0.22	-0.04
	52	58.00	0.04	-0.02	0.02	0.00	-0.02
►	53	97.00	-0.04	-0.02	-0.05	-0.02	0.08
	54	85.00	0.25	0.42	0.42	-0.26	-0.32
	55	73.00	0.37	0.34	-0.23	0.34	-0.24
►	56	10.00	-0.04	0.03	0.00	0.03	-0.03
	57	79.00	0.16	0.27	-0.05	0.27	-0.31
	58	93.00	0.16	0.29	0.29	-0.28	-0.14
	59	88.00	0.16	0.36	0.36	-0.20	-0.32
	60	93.00	0.12	0.10	-0.11	-0.02	0.10
	61	88.00	0.25	0.31	0.00	0.31	-0.31
	62	62.00	0.58	0.47	-0.23	-0.34	0.47

**Figure 5 *TAD* Sample Output using EI, DI and Point Biserial (Part 2 of 2)<sup>2</sup>**

## **ITEMAN Sample Output Using EI, DI, and Point Biserial Correlations**

The sample output below is from the *ITEMAN* software program using a sample data file provided by with the demonstration program. The data file consists of 400 students' test scores from a 20-item test. The demonstration program will not work with other data files other than the ones provided with the program. The Ease Index, Differentiation Index, and Point Biserial Correlation Coefficient values are displayed.

### **Item analysis for data from file C:\ITEMAN\SAMPLE1.DAT**

Date: 01/01/01

Time: 2:08 PM

\*\*\* NOTE \*\*\* This demonstration version of the program can be used only with the sample data provided. Other uses may result in incorrect item, alternative, and scale statistics and in incorrect examinee scores.

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics					
		Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	.79	.37	.37	1	.16	.29	.04	-.24	
					2	.04	.10	.00	-.23	
					3	.01	.03	.00	-.21	
					4	.79	.59	.96	.37	*
					Other	.00	.00	.00		
2	0-2	.69	.69	.60	1	.04	.10	.00	-.26	
					2	.69	.29	.97	.60	*
					3	.05	.15	.00	-.31	
					4	.22	.46	.03	-.38	
					Other	.00	.00	.00		
3	0-3	.94	.18	.49	1	.04	.12	.00	-.39	
					2	.01	.03	.00	-.18	
					3	.94	.82	1.00	.49	*
					4	.01	.04	.00	-.23	
					Other	.00	.00	.00		

**Figure 6 *ITEMAN* Sample Output Using EI, DI and Point Biserial (Part 1 of 5)<sup>3</sup>**

4	0-4	.91	.22	.37	1	.91	.77	.99	.37	*
					2	.00	.00	.00		
					3	.08	.18	.01	-.30	
					4	.01	.04	.00	-.23	
					Other	.00	.00	.00		
5	0-5	.97	.03	.13	1	.02	.03	.01	-.14	*
					2	.00	.01	.00	-.03	
					3	.00	.00	.00	.01	
					4	.97	.96	.99	.13	
					Other	.00	.00	.00		
6	0-6	.70	.55	.51	1	.17	.32	.03	-.34	*
					2	.70	.40	.95	.51	
					3	.07	.16	.01	-.25	
					4	.06	.12	.01	-.18	
					Other	.00	.00	.00		
7	0-7	.94	.08	.07	1	.01	.03	.00	-.05	*
					2	.94	.91	.99	.07	
					3	.01	.03	.00	-.09	
					4	.04	.03	.01	-.01	
					Other	.00	.00	.00		
8	0-8	.85	.26	.33	1	.03	.08	.00	-.29	*
					2	.06	.13	.02	-.21	
					3	.06	.08	.02	-.08	
					4	.85	.71	.97	.33	
					Other	.00	.00	.00		
9	0-9	.85	.35	.47	1	.01	.03	.00	-.15	*
					2	.85	.62	.97	.47	
					3	.09	.21	.02	-.31	
					4	.05	.14	.01	-.29	
					Other	.00	.00	.00		
10	0-10	.65	.61	.49	1	.65	.34	.95	.49	*
					2	.03	.07	.00	-.24	
					3	.04	.10	.02	-.13	
					4	.28	.50	.03	-.38	
					Other	.00	.00	.00		

**Figure 7 ITEMAN Sample Output Using EI, DI and Point Biserial (Part 2 of 5)<sup>4</sup>**

11	0-11	.56	.69	.51	1	.56	.24	.93	.51	*
					2	.09	.17	.01	-.19	
					3	.28	.47	.05	-.33	
					4	.06	.12	.01	-.20	
					Other	.00	.00	.00		
12	0-12	.97	.08	.29	1	.01	.02	.00	-.11	*
					2	.02	.07	.00	-.27	
					3	.97	.92	1.00	.29	
					4	.00	.00	.00		
					Other	.00	.00	.00		
13	0-13	.93	.18	.33	1	.06	.15	.00	-.27	*
					2	.00	.02	.00	-.15	
					3	.00	.02	.00	-.15	
					4	.93	.82	1.00	.33	
					Other	.00	.00	.00		
14	0-14	.87	.35	.43	1	.02	.05	.00	-.13	*
					2	.87	.65	1.00	.43	
					3	.04	.12	.00	-.32	
					4	.08	.18	.00	-.26	
					Other	.00	.00	.00		
15	0-15	.93	.16	.36	1	.01	.03	.00	-.15	*
					2	.01	.01	.00	-.03	
					3	.05	.13	.00	-.34	
					4	.93	.84	1.00	.36	
					Other	.00	.00	.00		
16	0-16	.95	.15	.40	1	.95	.85	1.00	.40	*
					2	.00	.02	.00	-.13	
					3	.03	.08	.00	-.29	
					4	.02	.06	.00	-.23	
					Other	.00	.00	.00		
17	0-17	.96	.12	.42	1	.01	.04	.00	-.27	*
					2	.02	.08	.00	-.32	
					3	.00	.00	.00		
					4	.96	.88	1.00	.42	
					Other	.00	.00	.00		

**Figure 8 *ITEMAN* Sample Output Using EI, DI and Point Biserial (Part 3 of 5)<sup>5</sup>**



18	0-18	.82	.34	.37	1	.82	.64	.98	.37	*
					2	.05	.11	.01	-.22	
					3	.05	.09	.00	-.18	
					4	.08	.16	.01	-.21	
					Other	.00	.00	.00		
19	0-19	.84	.28	.36	1	.13	.24	.03	-.29	*
					2	.02	.03	.00	-.12	
					3	.84	.69	.97	.36	
					4	.01	.04	.00	-.15	
					Other	.00	.00	.00		
20	0-20	.48	.67	.47	1	.43	.61	.10	-.33	*
					2	.48	.23	.90	.47	
					3	.05	.08	.00	-.11	
					4	.05	.09	.00	-.24	
					Other	.00	.00	.00		

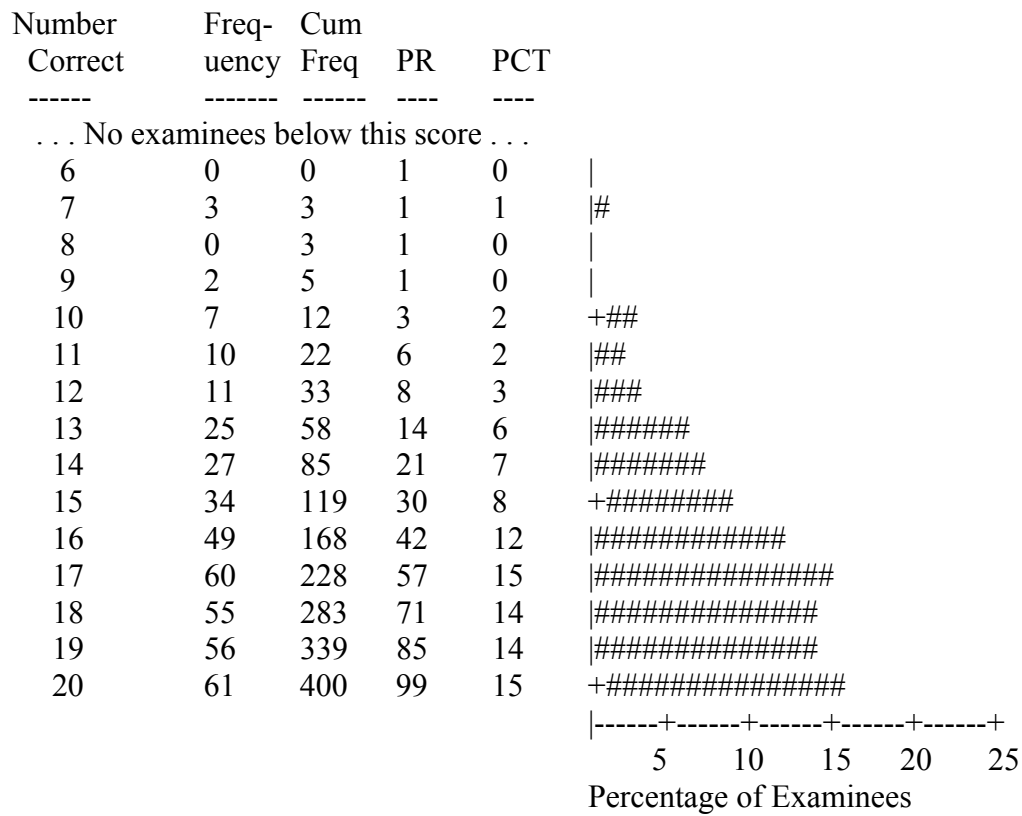
There were 400 examinees in the data file.

#### Scale Statistics

-----

N of Items        20  
 N of Examinees   400  
 Mean            16.605  
 Variance        7.499  
 Std. Dev.       2.738  
 Skew            -0.838  
 Kurtosis        0.389  
 Minimum        7.000  
 Maximum        20.000  
 Median          17.000  
 Alpha            0.712  
 SEM            1.470  
 Mean P          0.830  
 Mean Item-Tot.   0.389  
 Mean Biserial    0.641  
 Max Score (Low)   15  
 N (Low Group)    119  
 Min Score (High)   19  
 N (High Group)   117

**Figure 9 *ITEMAN* Sample Output Using EI, DI and Point Biserial (Part 4 of 5)<sup>6</sup>**



**Figure 10** *ITEMAN* Sample Output Using EI, DI and Point Biserial (Part 5 of 5)<sup>7</sup>

## **ITEMAN Sample Output Using EI, DI, and Biserial Correlations**

The sample output below is from the *ITEMAN* software program using the same sample data file as used in the previous figures. The only difference is that this sample output displays Biserial Coefficients instead of Point Biserial Coefficients. Again, the heading “Prop. Correct” is the same as Ease Index and the “Disc. Index” is the same as the Differentiation Index.

### **Item analysis for data from file C:\ITEMAN\SAMPLE1.DAT**

Date: 01/01/01

Time: 3:15 PM

\*\*\* NOTE \*\*\* This demonstration version of the program can be used only with the sample data provided. Other uses may result in incorrect item, alternative, and scale statistics and in incorrect examinee scores.

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics					
		Prop. Correct	Disc. Index	Biser.	Alt.	Prop. Total	Endorsing Low	High	Biser.	Key
1	0-1	.79	.37	.53	1	.16	.29	.04	-.36	
					2	.04	.10	.00	-.52	
					3	.01	.03	.00	-.87	
					4	.79	.59	.96	.53	*
					Other	.00	.00	.00		
2	0-2	.69	.69	.79	1	.04	.10	.00	-.61	
					2	.69	.29	.97	.79	*
					3	.05	.15	.00	-.63	
					4	.22	.46	.03	-.54	
					Other	.00	.00	.00		
3	0-3	.94	.18	1.00	1	.04	.12	.00	-.93	
					2	.01	.03	.00	-.74	
					3	.94	.82	1.00	1.00	*
					4	.01	.04	.00	-.79	
					Other	.00	.00	.00		
4	0-4	.91	.22	.65	1	.91	.77	.99	.65	*
					2	.00	.00	.00		

**Figure 11 *ITEMAN* Sample Output Using EI, DI and Biserial (Part 1 of 5)<sup>8</sup>**

					3	.08	.18	.01	-.55	
					4	.01	.04	.00	-.79	
					Other	.00	.00	.00		
5	0-5	.97	.03	.33	1	.02	.03	.01	-.39	
					2	.00	.01	.00	-.14	
					3	.00	.00	.00	.05	
					4	.97	.96	.99	.33	*
					Other	.00	.00	.00		
6	0-6	.70	.55	.68	1	.17	.32	.03	-.51	
					2	.70	.40	.95	.68	*
					3	.07	.16	.01	-.47	
					4	.06	.12	.01	-.36	
					Other	.00	.00	.00		
7	0-7	.94	.08	.13	1	.01	.03	.00	-.16	
					2	.94	.91	.99	.13	*
					3	.01	.03	.00	-.32	
					4	.04	.03	.01	-.02	
					Other	.00	.00	.00		
8	0-8	.85	.26	.50	1	.03	.08	.00	-.73	
					2	.06	.13	.02	-.41	
					3	.06	.08	.02	-.15	
					4	.85	.71	.97	.50	*
					Other	.00	.00	.00		
9	0-9	.85	.35	.72	1	.01	.03	.00	-.56	
					2	.85	.62	.97	.72	*
					3	.09	.21	.02	-.56	
					4	.05	.14	.01	-.59	
					Other	.00	.00	.00		
10	0-10	.65	.61	.63	1	.65	.34	.95	.63	*
					2	.03	.07	.00	-.64	
					3	.04	.10	.02	-.30	
					4	.28	.50	.03	-.50	
					Other	.00	.00	.00		

**Figure 12 *ITEMAN* Sample Output Using EI, DI and Biserial (Part 2 of 5)<sup>9</sup>**

11	0-11	.56	.69	.64	1	.56	.24	.93	.64	*
----	------	-----	-----	-----	---	-----	-----	-----	-----	---

					2	.09	.17	.01	-.34	
					3	.28	.47	.05	-.44	
					4	.06	.12	.01	-.38	
					Other	.00	.00	.00		
12	0-12	.97	.08	.73	1	.01	.02	.00	-.48	
					2	.02	.07	.00	-.74	
					3	.97	.92	1.00	.73	*
					4	.00	.00	.00		
					Other	.00	.00	.00		
13	0-13	.93	.18	.63	1	.06	.15	.00	-.54	
					2	.00	.02	.00	-.71	
					3	.00	.02	.00	-.71	
					4	.93	.82	1.00	.63	*
					Other	.00	.00	.00		
14	0-14	.87	.35	.67	1	.02	.05	.00	-.39	
					2	.87	.65	1.00	.67	*
					3	.04	.12	.00	-.75	
					4	.08	.18	.00	-.47	
					Other	.00	.00	.00		
15	0-15	.93	.16	.69	1	.01	.03	.00	-.56	
					2	.01	.01	.00	-.12	
					3	.05	.13	.00	-.73	
					4	.93	.84	1.00	.69	*
					Other	.00	.00	.00		
16	0-16	.95	.15	.83	1	.95	.85	1.00	0.83	*
					2	.00	.02	.00	-.64	
					3	.03	.08	.00	-.73	
					4	.02	.06	.00	-.70	
					Other	.00	.00	.00		
17	0-17	.96	.12	1.00	1	.01	.04	.00	-.93	
					2	.02	.08	.00	-.90	
					3	.00	.00	.00		
					4	.96	.88	1.00	1.00	*
					Other	.00	.00	.00		

**Figure 13 *ITEMAN* Sample Output Using EI, DI and Biserial (Part 3 of 5)<sup>10</sup>**

18	0-18	.82	.34	.55	1	.82	.64	.98	.55	*
					2	.05	.11	.01	-.45	
					3	.05	.09	.00	-.39	
					4	.08	.16	.01	-.38	
					Other	.00	.00	.00		
19	0-19	.84	.28	.54	1	.13	.24	.03	-.47	*
					2	.02	.03	.00	-.36	
					3	.84	.69	.97	.54	
					4	.01	.04	.00	-.51	
					Other	.00	.00	.00		
20	0-20	.48	.67	.59	1	.43	.61	.10	-.42	*
					2	.48	.23	.90	.59	
					3	.05	.08	.00	-.23	
					4	.05	.09	.00	-.52	
					Other	.00	.00	.00		

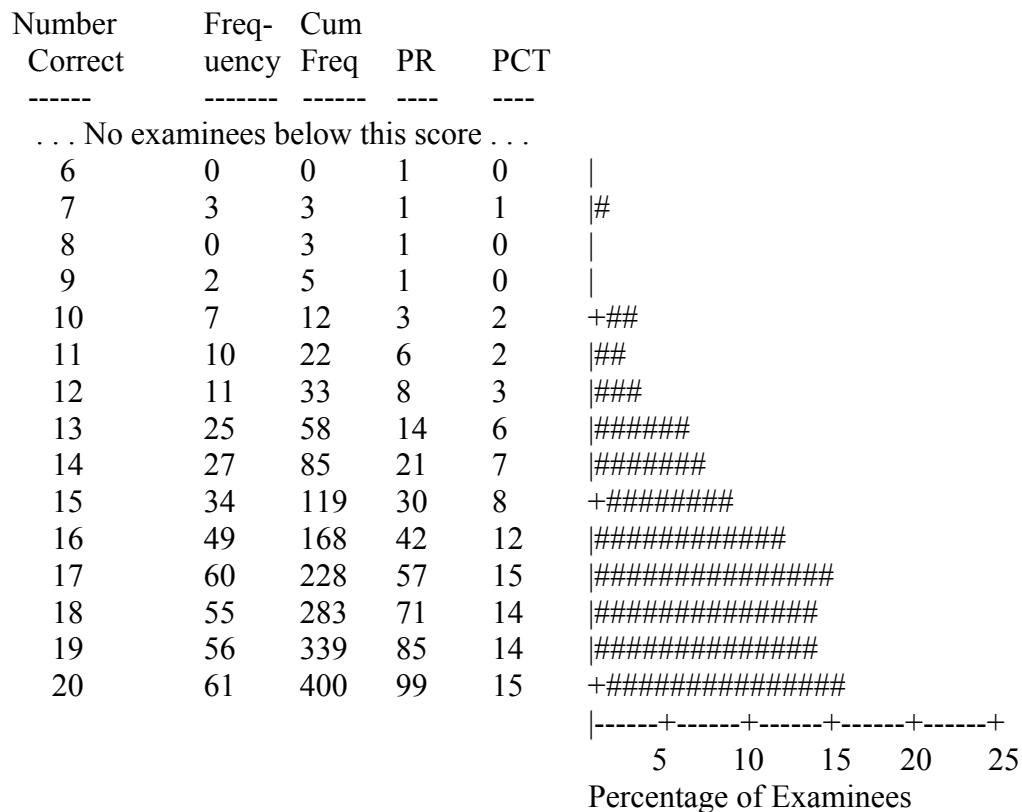
There were 400 examinees in the data file.

#### Scale Statistics

-----

N of Items        20  
 N of Examinees   400  
 Mean            16.605  
 Variance        7.499  
 Std. Dev.       2.738  
 Skew            -0.838  
 Kurtosis        0.389  
 Minimum        7.000  
 Maximum        20.000  
 Median          17.000  
 Alpha            0.712  
 SEM            1.470  
 Mean P          0.830  
 Mean Item-Tot.   0.389  
 Mean Biserial    0.641  
 Max Score (Low)   15  
 N (Low Group)    119  
 Min Score (High)   19  
 N (High Group)   117

**Figure 14 *ITEMAN* Sample Output Using EI, DI and Biserial (Part 4 of 5)<sup>11</sup>**



**Figure 15** *ITEMAN* Sample Output Using EI, DI and Biserial (Part 5 of 5)<sup>12</sup>

### Notes

<sup>1</sup> Output used with permission, Thomas R. Renckly, Jan. 12, 2001. Output from *Test Analysis & Development Sysem (TAD)* version 5.49. CD-ROM.(1990-2000).

<sup>2</sup> Ibid.

<sup>3</sup> Output used with permission, David J. Weiss, President, Assessment Systems Corporation, Jan. 11, 2001. Output from *ITEMAN* software program, demonstration version 3.50 Assessment Systems Corporation. (1995). <http://www.assess.com>.

<sup>4</sup> Ibid.

<sup>5</sup> Ibid.

<sup>6</sup> Ibid.

<sup>7</sup> Ibid.

<sup>8</sup> Ibid.

<sup>9</sup> Ibid.

<sup>10</sup> Ibid.

<sup>11</sup> Ibid.

<sup>12</sup> Ibid.

## Appendix B

### Question Writing Guidelines

This section provides both the Maxwell Academic Instructor School *Test Item Analysis Handout's* section on Qualitative Analysis as well as James Hansen and Lee Dexter's *Item Writing Guidelines*.

#### **Maxwell Academic Instructor School *Test Item Analysis Handout's* section on Qualitative Analysis<sup>1</sup>**

1. Test Construction
  - a) Item validity
  - b) Stem presents a meaningful problem to be solved
  - c) Use simple and clear wording
  - d) Avoid clue words
  - e) Avoid grammatical give-a-ways
  - f) Use equal length alternatives
  - g) Highlight key words
  - h) Use plausible distractors
  - i) Put all common wording in the stem
  - j) Ensure one clearly best answer
  - k) Use Positively stated stems when possible



## **James D. Hansen and Lee Dexter's Item-writing Guidelines<sup>2</sup>**

- 1) **Present a single, clearly formulated problem in the stem of the item.** If more than one problem is given and the student fails the question, it is not possible to identify which problem caused the error.
- 2) **State the stem in simple, clear language.** Poorly written or complex questions may cause knowledgeable students to answer incorrectly. Avoid unnecessary statements in the stem and do not continue teaching on an exam.
- 3) **Put as much wording as possible into the stem.** It is inefficient to repeat words, and students will have less difficulty with shorter items.
- 4) **When possible, state the stem in positive form.** Asking a student to identify an incorrect alternative does not necessarily test whether the student knows the correct answer. Knowing what is true is generally a more important learning outcome than knowing what is not true. Negatively phrased items are often written, however, because they are easier to create. Positively stated items require the author to devise three distractors for a four-alternative question, but a negatively stated item requires that only one plausible alternative be devised--the answer.
- 5) **Emphasize (by using italics and/or boldface) negative wording whenever it is used in the stem.** Not emphasizing negative wording may cause such wording to be overlooked.
- 6) **Be certain that the intended answer is correct or clearly the best.** Test quality will be improved and arguments from students will be lessened.
- 7) **Alternatives should be grammatically consistent with the stem and parallel in form.** Violations of this guideline may provide clues to the correct answer or aid students in eliminating distractors that do not match.

8) **Avoid verbal clues that may eliminate a distractor or lead to the correct answer.** There are several forms of verbal clues:

(a) **Avoid similarity of wording in the stem and the correct answer.** Similar wording can make the correct response more attractive to students who do not know the answer.

(b) **The correct answer should not be more detailed or include more textbook language than the distractors.**

(c) **Avoid absolute terms in the distractors.** Test-wise students will eliminate distractors containing words like "all," "only," or "never," because such statements are usually false.

(d) **Avoid pairs of responses that are all-inclusive.** This structure allows students to eliminate other alternatives because the inclusive pair covers all possibilities. An uninformed student would have a 50% chance of guessing the correct answer.

(e) **Avoid responses that have the same meaning.** Students will eliminate those alternatives because there can be only one correct answer.

(f) **If alternatives consist of pairs of answers, avoid a structure that yields the correct answer--an intersection of repeated terms.** For example, say the correct answer is x and y. To discriminate between students who know only part of the answer, an author might supply these alternatives:

x and y

x and z

w and y

l and p

A test-wise student who does not know the answer is attracted to the first alternative because the importance of x and y is signaled by their repetition.

- 9) **Make all distractors plausible to those who do not know the correct answer.** Good multiple-choice items depend on effective distractors.
- 10) **Avoid using "all of the above."** Students can select it as the correct answer by identifying any two alternatives as correct without knowing that they are all correct. Or, students can eliminate it by observing that any one alternative is wrong.
- 11) **Use "none of the above" with caution.** This may only measure the ability to detect incorrect answers. Although this alternative is more defensible in computation-type problems, "none of the above" is often used where the author has difficulty devising another plausible distractor.
- 12) **Follow the normal rules of grammar and punctuation.** For example, stems in question form should have alternatives that begin with capital letters. Alternatives in statement completion items should begin with lower-case letters. Periods should not be used with numerical alternatives, to avoid confusion with decimal points.

### Notes

<sup>1</sup> *Test Item Analysis Handout*, AI-641-c, 02-00, Academic Instructor School, Maxwell AFB, AL, p. 5.

<sup>2</sup> Guidelines used with permission, Mary J. Winolm, Copyright Officer, Heldref Publications. Guidelines from James D. Hansen and Lee Dexter's article "Quality Multiple-Choice Test Questions: Item-Writing Guidelines and an Analysis of Auditing Testbanks," *Journal for Education for Business*, vol. 73, no. 2, (Nov 1997), pp. 95-6 Heldref Publications.

## ***Glossary***

### **Abbreviations**

ACSC	Air Command and Staff College
ACT	American College Testing
AU	Air University
AWC	Air War College
BIS	Biserial Correlation Coefficient
CLEP	College-Level Examination Program
DI	Differentiation Index or Discrimination Index
DoD	Department of Defense
EI	Ease Index
<i>ITEMAN</i>	<i>ITEMAN</i> software program
N-O-T-A	“none-of-the-above” alternative (answer)
Rpbis	Point Biserial Correlation Coefficient
SAT	Scholastic Assessment Test
<i>TAD</i>	<i>Test Analysis &amp; Development System</i> software program
USAF	United States Air Force

### **Definitions**

**alternative.** One of the answers of a multiple-choice question.

**Biserial Correlation Coefficient (BIS).** This statistic correlates overall test scores to the correct answering of an individual test item (question).<sup>1</sup> In other words, it is a measurement of how getting a particular question correct correlates to a high score (or passing grade) on the test. Possible values range from –1.0 to 1.0. A score of –1.0 would indicate that all those who answered the question correctly scored poorly on (or failed) the test. A score of 1.0 would indicate that those who answered the question correctly scored well on (or passed) the test.

**Brennan's B Coefficient.** *From the TAD software help manual:* "This coefficient is computed during item analysis when the user identifies a mastery criterion group in the student population being analyzed. You may identify students as Masters or Nonmasters either automatically by their test scores in relation to the passing cutoff score, or you may select them manually from a list displayed on-screen during test analysis. Once Masters and Nonmasters are identified in the student population, the B coefficient can be calculated for each test question as the difference between the number of students in the Masters (upper scoring) group who answered the test question correctly and the number of students in the Nonmasters (lower scoring) group who answered it correctly.

As with the norm-referenced discrimination index, a positive B coefficient indicates that a larger proportion of the upper group answered the question correctly than the lower group; a negative value indicates just the opposite. Thus, the B coefficient may be viewed as a coefficient of discrimination in terms of the test question's ability to discriminate between Masters' and Nonmasters' ability or knowledge levels.

Since discrimination is not necessarily the goal of criterion-referenced instruction, a value of zero for the B coefficient is considered ideal from a criterion-referenced perspective."<sup>2</sup>

**Differentiation Index (DI).** This statistic is a measure of each test question's ability to differentiate between high scoring and low scoring students. This is computed as: the number of people with highest test scores (top 27%) who answered the item correctly minus the number of people with lowest scores (bottom 27%) who answered the item correctly, divided by the number of people in the largest of the two groups.<sup>3</sup> The higher the number, the more the question is able to discriminate the higher scoring people from the lower scoring people. Possible values range from -1.0 to 1.0. Sometimes 25% or 33.3% is used instead of 27% (for the top and bottom group of test scores).

*From the TAD software help manual:* "This statistic, identified as DI in the TAD program, is computed by first rank-ordering students from highest to lowest test score. Next, the student group is divided into third in such a way that the upper and lower thirds are always kept equal. Then, for each test question, the DI is computed as the number of students in the high third answering the question correctly minus the number of students in the low third answering the question correctly, divided by one-third of the total number of students taking the test. This statistic measures each test question's ability to differentiate between high-and-low-achieving students.

The DI can range in value from -1.0 to +1.0. A positive DI indicates that more high-achieving students answered the question correctly than low-achieving students. When the number of low-achieving students answering a question correctly becomes greater than the number of high-achieving students, the DI becomes negative, and signals a possible problem area.

The DI only makes use of two-thirds of the available test scores (the middle-third is not used). Also, when tie scores occur at either the upper- or lower-third cutoff points in the score distribution, there is not reliable method of selecting which of the ties are placed into these (countable) groups and which are placed in the (unused) middle group. This is problematic since the student whose score was placed into the middle group may have positively or negatively affected the DI computation had it been included into the upper- or lower-third group. Since these tie-score placements are typically done randomly, the same

test question could yield different DI values on different computations depending on which students were included into the countable groups.”<sup>4</sup>

**distractor.** An incorrect answer of a multiple-choice question.

**Ease Index (EI).** This is also known as Difficulty Index, Item Difficulty, Percent Correct or “p-value”. It is simply the proportion (or percentage) of students taking the test who answered the item correctly.<sup>5</sup> This value is usually reported as a proportion (rather than percentage), ranging from 0.0 to 1.0. A value of 0.0 indicates that no one answered the item correctly. A value of 1.0 indicates that everyone answered the item correctly. A value of 0.5 indicates that half the class answered correctly.

*From the TAD software help manual:* “This statistic is sometimes called the Difficulty Index in some psychometric texts. The name Ease Index (or EI as it is referred to in the TAD program) is a more appropriate name for this statistic because higher indexes relate to easier test questions while lower indexes relate to more difficult questions. It is computed for each test question by dividing the number of students who answered the question correctly by the total number of students taking the test, and multiplying the result by 100 to produce a percentage. Higher Ease Indexes relate to test questions that are answered correctly by a larger proportion of students; thus, the easier the question appears to be. This statistic can range in value from 0 to 100. Values above about 70 indicate an especially easy test question, while values below about 45 indicate a relatively difficult test question.”<sup>6</sup>

**item.** A multiple-choice question.

**option.** An answer of a multiple-choice question. Same as **alternative**.

**Point Biserial Correlation Coefficient (Rpbis).** This statistic is a measure of the capacity of a test item (question) to discriminate between high and low scores.<sup>7</sup> In other words, it is how much predictive power an item has on overall test performance. Possible values range from -1.0 to 1.0 (the maximum value can never reach 1.0, and the minimum can never reach -1.0). A value of 0.6 would indicate the question has a good predictive power, i.e., those who answered the item correctly received a higher average grade compared to those who answered the item incorrectly. A value of -0.6 would indicate the question has a poor predictive power, i.e., those who answered the item incorrectly received a higher average grade compared to those who answered the item correctly.

*From the TAD software help manual:* “This statistic is computed as an alternative to the more typical (but less stable) differentiation index (DI). It can be interpreted in exactly the same way as the DI. The higher stability of the Point Biserial Coefficient derives from two facts: (1) this coefficient makes use of all test data, and (2) the computation does not depend on arbitrary cutoff values (as does the DI). Both the Point Biserial Coefficient and the DI are computed and displayed by the program. You may use whichever statistic you are accustomed to. However, for those users who are not particularly disposed toward one or the other statistic, we recommend using the Point Biserial Correlation Coefficient. You will find when comparing both statistics side-by-side that if the DI indicates a value of 0, the Point Biserial Coefficient may very often indicate a non-zero value. This reflects the additional information provided by the middle third of the student group, which is not contained in the DI.”<sup>8</sup>

**threshold value.** It is the lowest value that will not flag a question. A value lower than the threshold value will flag the question.

## Notes

<sup>1</sup> Ferguson, George A. (1976). *Statistical Analysis in Psychology and Education* (4<sup>th</sup> ed.), New York, NY, McGraw-Hill, p. 418.

<sup>2</sup> Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

<sup>3</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>, p.6

<sup>4</sup> Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

<sup>5</sup> Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>. p.1.

<sup>6</sup> Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

<sup>7</sup> Ferguson, George A. (1976). *Statistical Analysis in Psychology and Education* (4<sup>th</sup> ed.), New York, NY, McGraw-Hill, p. 416.

<sup>8</sup> Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

## ***Bibliography***

- Attali, Yigal and Fraenkel, Tamar. "The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative," *Journal of Educational Measurement*, vol. 37, no. 1, (Spring 2000), 77-86.
- Ballantyne, Christina. (2000). "Multiple-Choice Tests: Test Scoring and Analysis" <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html> .
- Brown, Frederick G. (1983). *Principles of Educational and Psychological Testing* (3<sup>rd</sup> ed.), New York, NY, Holt, Rinehart and Winston.
- Cizek, Gregory J., O'Day, Denis M. "Further Investigation of Nonfunctioning Options in Multiple-Choice Test Items," *Educational & Psychological Measurement*, vol. 54, issue 4, (Winter 1994), 861-872.
- Copeland, Karen, PhD. (2000). "Perusing Process Data with JMP Histograms," <http://www.sas.com/service/library/periodicals/obs/obswww17/index.html> .
- Dawson, Lee D. "Statistics With a Lot Extras," *Quality Digest*. May 98.
- Dean, James and Parmigiani, Rosemary. "Multimedia Reviews: CD-ROM," *Media and Methods*, vol. 36, issue 5, (May/June 2000), 19-22.
- Ferguson, George A. (1976). *Statistical Analysis in Psychology and Education* (4<sup>th</sup> ed.), New York, NY, McGraw-Hill.
- Garrett, Henry E. (1966). *Statistics in Psychology and Education* (6<sup>th</sup> ed.), New York, NY, David McKay Co.
- Geiger, Marshall A. and Simmons, Kathleen A. "Intertopical Sequencing of Multiple-Choice Questions: Effect on Exam Performance and Testing Time" *Journal of Education & Business*, vol. 70, issue 2, (Nov/Dec 1995), 87-90.
- Gronlund, Norman E. and Linn, Robert L. (1990). *Measurement and Evaluation in Teaching* (6<sup>th</sup> ed.), New York, NY, MacMilan.
- Haladyna, Thomas M. and Downing Steven M. "How Many Options is Enough for a Multiple-Choice Test Item?" *Educational & Psychological Measurement*, vol. 53, issue 4, (Winter 1993), 999-1010.
- Hansen, James D. and Dexter, Lee. "Quality Multiple-Choice Test Questions: Item-Writing Guidelines and an Analysis of Auditing Testbanks," *Journal for Education for Business*, vol. 73, no. 2, (Nov 1997), 94-97, Heldref Publications.
- ITEMAN software program, demonstration version 3.50 Assessment Systems Corporation. (1995). <http://www.assess.com> .
- ITEMAN Online Manual. Assessment Systems Corporation. <http://www.assess.com> .
- Kehoe, Jerard. (1995). "Basic Item Analysis for Multiple-Choice Tests," <http://ericae.net/digests/tm9511.htm> .
- Knowles, Susan L. and Welch, Cynthia A. "A Meta-Analytic Review of Item Discrimination and Difficulty in Multiple-Choice Items Using "None-of-the-Above"," *Educational & Psychological Measurement*, vol. 52, issue 3, (Fall 1992), 571-577.



Linn, Robert L. and Gronlund, Norman E. (1995). *Measurement and Assessment in Teaching*, (7<sup>th</sup> ed.), New York, NY, Merrill.

Matlock-Hetzel, Susan. (1997). "Basic Concepts in Item and Test Analysis", Texas A&M University. <http://cleo.murdoch.edu.au/evaluations/pubs/mcq/scpre.html>

Payne, David A. (1967). *Educational and Psychological Measurement: Contributions to Theory and Practice*, Waltham, MA, Blaisdell.

Renckly, Thomas R. *Test Analysis & Development Sysem (TAD) version 5.49*. CD-ROM.(1990-2000).

*Test Item Analysis Handout*, AI-641-c, 02-00, Academic Instructor School, Maxwell AFB, AL.

Zieky, Michael. Email concerning ETS's item analysis, 06 Nov. 2000. Email address: [mzieky@ets.org](mailto:mzieky@ets.org).